METHOD AND SYSTEM FOR MONITORING EVENTS IN STORAGE AREA NETWORKS

INVENTOR(S):

EDWARD C. MCGLAUGHLIN

JOHN P. WAGNER

WILLIAM C. BACKSTROM

BACKGROUND

[0001] Field of the Invention

[0002] The present invention relates to storage area networks, and more particularly, to monitoring events in storage area networks.

[0003] 2. Background of the Invention

[0004] Storage Area Networks ("SANs") provide multiple paths to host computing systems to access memory devices (or storage devices). The multiple paths allow host systems to access data in the event of a failure.

[0005] Various standards are used for operating SANs. One such standard is Fibre Channel. Fibre channel is a set of American National Standard Institute (ANSI) standards, which provide a serial transmission protocol for storage and network protocols such as HIPPI, SCSI, IP, ATM and others. Fibre channel provides an input/output interface

to meet the requirements of both channel and network users.

[0006] Fibre channel supports three different topologies: point-to-point, arbitrated loop and fibre channel fabric. The point-to-point topology attaches two devices directly. The arbitrated loop topology attaches devices in a loop. The fibre channel fabric topology attaches host systems directly to a fabric, which are then connected to multiple devices. The fibre channel fabric topology allows several media types to be interconnected.

[0007] Fibre channel is a closed system that relies on multiple ports to exchange information on attributes and characteristics to determine if the ports can operate together. If the ports can work together, they define the criteria under which they communicate.

[0008] In fibre channel, a path is established between two nodes where the path's primary task is to transport data from one point to another at high speed with low latency, performing only simple error detection in hardware.

[0009] Fibre channel fabric devices include a node port or "N_Port" that manages fabric connections.

The N_port establishes a connection to a fabric element (e.g., a switch) having a fabric port or F_port. Fabric elements include the intelligence to handle routing, error detection, recovery, and similar management functions.

[00010] A fibre channel switch is a multi-port device where each port manages a simple point-to-point connection between itself and its attached system. Each port can be attached to a server, peripheral, I/O subsystem, bridge, hub, router, or even another switch. A switch receives messages from one port and automatically routes it to another port. Multiple calls or data transfers happen concurrently through the multi-port fibre channel switch.

[00011] Fibre channel switches use memory buffers to hold frames received and sent across a network. Associated with these buffers are credits, which are the number of frames a Fibre Channel port can transmit without overflowing the receive buffers at the other end of the link. Receiving an R_RDY primitive signal increases the credit, and sending a frame decreases the credit.

[00012] In conventional SAN systems, if an input/output operation is "timed out", (i.e. it

takes longer than say a threshold value) it results in an application time-out. This results in a prolonged recovery operation and reduces the overall availability of data to a user using a host computing system. Hence, in conventional systems, recovery occurs after a failure has already occurred.

[00013] As SANs become more complex with higher bandwidth requirements there is a need for detecting failure and re-routing data requests before the actual failure occurs and disrupts information flow within the SAN.

[00014] SUMMARY OF THE INVENTION

[00015] In one embodiment of the present invention, a system for monitoring at least one event and detecting at least one indicator that precedes a failure in a storage area network is provided. The system includes, a switch element having a port-monitoring module that detects when a fibre channel port link indicator varies from a threshold value for the fibre channel port link parameter.

[00016] The port-monitoring module sends a message to a performance-monitoring module to process an event when the fibre channel port link indicator

varies from the threshold value. The performance-monitoring module may notify a host computer when the fibre channel port link indicator varies from the threshold value, and the fibre channel port whose port link parameter varies from the threshold value may be taken off-line.

[00017] In yet another embodiment of the present invention, the system includes a fabric monitoring module that detects when a remote fabric switch and a local fabric switch cannot exchange information and then sends a message to a threshold monitoring module.

[00018] The host system is notified that the remote and the local fabric switch cannot exchange information.

[00019] In yet another embodiment of the present invention, the system includes a chassis monitoring module that detects when an indicator varies from a threshold value and notifies the threshold monitoring module of the variation.

[00020] In yet another embodiment of the present invention, the threshold-monitoring module receives a message from at least one monitoring agent indicating if an indicator varies from a threshold value, and in response to the message,

the threshold monitoring module coordinates an event. The threshold values are stored and updated in a threshold table. The monitoring agent includes the Chassis monitoring module, the fabric monitoring module, the port monitoring module, and/or an Nx_port event monitoring module.

[00021] The threshold-monitoring module notifies an event response module when an indicator value varies from a threshold value and an event is generated in response to the notification based on an event table.

[00022] In yet another embodiment of the present invention, a method for monitoring at least one event and detecting at least one indicator that precedes a failure in a storage area network is provided. The method includes, monitoring at least one event by using at least one monitoring agent in a fabric switch; comparing an indicator value to a threshold value for the indicator; and notifying a threshold-monitoring module if the parameter value varies from the threshold value.

[00023] In one aspect of the present invention, various events are monitored in real time, which can result in the failure of certain SAN components/services. This allows a system to be

intelligent and preemptive, which avoids disruption of SANs.

[00024] This brief summary has been provided so that the nature of the invention may be understood quickly. A more complete understanding of the invention can be obtained by reference to the following detailed description of the preferred embodiments thereof concerning the attached drawings.

[00025] BRIEF DESCRIPTION OF THE DRAWINGS

[00026] The foregoing features and other features of the present invention will now be described with reference to the drawings of a preferred embodiment. In the drawings, the same components have the same reference numerals. The illustrated embodiment is intended to illustrate, but not to limit the invention. The drawings include the following Figures:

[00027] Figure 1A is a block diagram of a fibre channel network system;

[00028] Figures 1B-1D show block diagrams of various switch element configurations used according to one aspect of the present invention;

[00029] Figure 2 shows a block diagram of a system for monitoring events and detecting failure, according to one aspect of the present invention; and

[00030] Figure 3 shows a process flow diagram for monitoring events, according to one aspect of the present invention.

[00031] DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[00032] Definitions:

[00033] The following definitions are provided as they are typically (but not exclusively) used in the fibre channel environment, implementing the various adaptive aspects of the present invention.

[00034] "E-Port": A fabric expansion port that attaches to another Interconnect port to create an Inter-Switch Link.

[00035] "F-Port": A port to which non-loop N_Ports are attached to a fabric and does not include FL_ports.

[00036] "Fibre channel ANSI Standard": The standard describes the physical interface, transmission and signaling protocol of a high performance serial link for support of other high level protocols associated with IPI, SCSI, IP, ATM and others.

[00037] "FC-1": Fibre channel transmission protocol, which includes serial encoding, decoding and error control.

[00038] "FC-2": Fibre channel signaling protocol that includes frame structure and byte sequences.

[00039] "FC-3": Defines a set of fibre channel services that are common across plural ports of a node.

[00040] "FC-4": Provides mapping between lower levels of fibre channel, IPI and SCSI command sets, HIPPI data framing, IP and other upper level protocols.

[00041] "Fabric": A system which interconnects various ports attached to it and is capable of routing fibre channel frames by using destination identifiers provided in FC-2 frame headers.

[00042] "Fabric Topology": This is a topology where a device is directly attached to a fibre channel fabric that uses destination identifiers embedded in frame headers to route frames through a fibre channel fabric to a desired destination.

[00043] "FL_Port": A L_Port that is able to perform the function of a F_Port, attached via a link to one or more NL_Ports in an Arbitrated Loop topology.

[00044] "Inter-Switch Link": A Link directly connecting the E_port of one switch to the E_port of another switch.

[00045] Port: A general reference to N. Sub.-- Port or F.Sub.--Port.

[00046] "L_Port": A port that contains Arbitrated Loop functions associated with the Arbitrated Loop topology.

[00047] "N-Port": A direct fabric attached port.

[00048] "NL_Port": A L_Port that can perform the function of a N_Port.

[00049] "Switch": A fabric element conforming to the Fibre Channel Switch standards.

[00050] To facilitate an understanding of the preferred embodiment, the general architecture and operation of a fibre channel system will be described. The specific architecture and operation of the preferred embodiment will then be described with reference to the general architecture of the fibre channel system.

[00051] Figure 1A is a block diagram of a fibre channel system 100 implementing the methods and systems in accordance with the adaptive aspects of the present invention. System 100 includes plural devices that are interconnected. Each device includes one or more ports, classified as node ports (N_Ports), fabric ports (F_Ports), and expansion ports (E_Ports). Node ports may be

located in a node device, e.g. server 103, disk array 105 and storage device 104. Fabric ports are located in fabric devices such as switch 101 and 102. Arbitrated loop 106 may be operationally coupled to switch 101 using arbitrated loop ports (FL_Ports).

[00052] The devices of Figure 1A are operationally coupled via "links" or "paths". A path may be established between two N_ports, e.g. between server 103 and storage 104. A frame-switched path may be established using multiple links, e.g. an N-Port in server 103 may establish a path with disk array 105 through switch 102.

[00053] Figure 1B is a block diagram of a 20-port ASIC fabric element according to one aspect of the present invention. Figure 1B provides the general architecture of a 20-channel switch chassis using the 20-port fabric element. Fabric element includes ASIC 20 with non-blocking fibre channel class 2 (connectionless, acknowledged) and class 3 (connectionless, unacknowledged) service between any ports. It is noteworthy that ASIC 20 may also be designed for class 1 (connection-oriented) service, within the scope and operation of the present invention as described herein.

[00054] The fabric element of the present invention is presently implemented as a single CMOS ASIC, and for this reason the term "fabric element" and ASIC are used interchangeably to refer to the preferred embodiments in this specification. Although Figure 1B shows 20 ports, the present invention is not limited to any particular number of ports.

[00055] ASIC 20 has 20 ports numbered in Figure 1B as GL0 through GL19. These ports are generic to common Fibre Channel port types, for example, F_Port, FL_Port and E-Port. In other words, depending upon what it is attached to, each GL port can function as any type of port. Also, the GL port may function as a special port useful in fabric element linking, as described below.

[00056] For illustration purposes only, all GL ports are drawn on the same side of ASIC 20 in Figure 1B. However, the ports may be located on both sides of ASIC 20 as shown in other figures. This does not imply any difference in port or ASIC design. Actual physical layout of the ports will depend on the physical layout of the ASIC.

[00057] Each port GL0-GL19 has transmit and receive connections to switch crossbar 50. One connection

is through receive buffer 52, which functions to receive and temporarily hold a frame during a routing operation. The other connection is through a transmit buffer 54.

[00058] Switch crossbar 50 includes a number of switch crossbars for handling specific types of data and data flow control information. For illustration purposes only, switch crossbar 50 is shown as a single crossbar. Switch crossbar 50 is a connectionless crossbar (packet switch) of known conventional design, sized to connect 21 x 21 paths. This is to accommodate 20 GL ports plus a port for connection to a fabric controller, which may be external to ASIC 20.

[00059] In the preferred embodiments of switch chasis described herein, the fabric controller is a firmware-programmed microprocessor, also referred to as the input/out processor "IOP"). IOP 66 is shown in Figure 1C as a part of a switch chasis utilizing one or more of ASIC 20. As seen in Figure 1B, bi-directional connection to IOP 66 is routed through port 67, which connects internally to a control bus 60. Transmit buffer 56, receive buffer 58, control register 62 and Status register 64 connect to bus 60. Transmit buffer 56 and

receive buffer 58 connect the internal connectionless switch crossbar 50 to IOP 66 so that it can source or sink frames.

[00060] Control register 62 receives and holds control information from IOP 66, so that IOP 66 can change characteristics or operating configuration of ASIC 20 by placing certain control words in register 62. IOP 66 can read status of ASIC 20 by monitoring various codes that are placed in status register 64 by monitoring modules, as discussed below with respect to Figure 2.

[00061] Figure 1C shows a 20-channel switch chassis S2 using ASIC 20 and IOP 66. S2 will also include other elements, for example, a power supply (not shown). The 20 GL ports correspond to channel C0-C19. Each GL port has a serial/deserializer (SERDES) designated as S0-S19. Ideally, the SERDES functions are implemented on ASIC 20 for efficiency, but may alternatively be external to each GL port.

[00062] Each GL port has an optical-electric converter, designated as OE0-OE19 connected with its SERDES through serial lines, for providing fibre optic input/output connections, as is well

known in the high performance switch design. The converters connect to switch channels C0-C19. It is noteworthy that the ports can connect through copper paths or other means instead of optical-electric converters.

[00063] Figure 1D shows a block diagram of ASIC 20 with sixteen GL ports and four 10G port control modules designated as XG0-XG3 for four 10G ports designated as XGP0-XGP3. ASIC 20 include a control port 62A that is coupled to IOP 66 through a PCI connection 66A.

[00064] Figure 2 shows a block diagram of a system 200, according to one embodiment of the present invention that provides an I/O path guard mechanism that improves overall SAN system efficiency. System 200 includes a fabric switch 201, a host computer 202, a storage controller 203 and a remote fabric switch 204.

[00065] It is noteworthy that the present invention is not limited to any particular number of remote switches, host computers, fabric switches or number of ports. For example, although only one remote fabric switch 204 is shown to illustrate the adaptive aspects of the present invention, other remote switches can be used with system 200.

[00066] It is noteworthy that some of the components in modules 201-204 have not been shown, as they are well known in the art. For example, a host computer 202 includes a central processing unit, storage devices, random access memory, read-only memory, keyboard, video interfaces and other devices that have not been described.

[00067] Fabric switch 201 is of the type discussed above with respect to Figures 1B-1D. However, other configurations may be used with system 200. Fabric switch 201 has plural monitoring modules that monitor various indicators/parameters/events (may be referred to as an indicator, parameter or an event). Every indicator has a threshold and a variation from the threshold value can result in a failure. When the monitoring agent detects a variation, it notifies a performance threshold monitoring module, as described below, which processes the event depending upon the function of the indicator. A host computer may also be notified of a possible failure to take preemptive action.

[00068] Fabric switch 201 is coupled to a host computer 202. I/O Port 205 and 221 provide connectivity between host computer 202 and fabric

switch 201. It is noteworthy that plural I/O ports 205 and 221 may be used to couple plural host computing systems.

[00069] Fabric Switch 201 is also coupled to storage controller 203 using ports 210 and 225. Once again, plural number of ports 210 and 225 may be used to couple fabric switch 201 to more than one storage controller 203. In one aspect of the present invention, storage controller 203 is a redundant array of independent disks ("RAID") controller and/or host bus adapter ("HBA").

[00070] RAID controllers control access to plural disks. An HBA is an I/O adapter that sits between a host computer's bus and the fibre channel loop and manages the transfer of information between the two channels. Both the RAID controller and HBAs have various components, for example, a processor, memory, and arbitration modules that allow them to operate with host computers, storage systems, and fabric switches. The present invention is not limited to any particular structure or type of storage controller 203.

[00071] Turning now in more detail to fabric switch 201, it includes a Performance Threshold Monitor module 209, which monitors the threshold status

and events related to plural monitoring agents (discussed below) and compares the monitored status with threshold values maintained in Threshold Table 214. The monitoring agents include Chassis Monitor 218, Nx_Port Event Monitor 212, Port Monitor 216, and Fabric Monitor 215, which are described below.

[00072] Chassis Monitor 218 tracks the temperature of the switch chassis (not shown), the power supply status, and fan (not shown) speed using sensors 219. When a status value exceeds the threshold configured in Threshold Table 214, a control signal (this term as used in the specification includes commands and/or message packets) is sent from Chassis Monitor 218 to Performance Threshold Monitor 209 for further processing of the event.

[00073] Performance threshold monitor 209 may send a signal to host computer 202 and/or remote fabric switch 204 when an operational parameter varies from a threshold value in table 214. Host computer 202 may perform a shut down of fabric switch 201 upon such a signal and may select an alternative path for subsequent I/O transfers.

[00074] Nx_Port Event Monitor 212 receives events generated by SAN targets such as Storage

Controller 203. These events include failed internal path notifications and other internal threshold errors. Nx_Port Event Monitor 212 supports a plurality of instances of I/O Port B1 modules 210 to monitor a plurality of Storage Controller 203 SAN targets.

[00075] When an event is received, Nx_Port Event Monitor 212 sends a control signal to Performance Threshold Monitor 209 for further processing of the event. Nx_Port Event Monitor 212 is also coupled to a diagnostic control module 228 that performs diagnostics based on certain events and threshold variations. The diagnostic information may be sent to host computer 202 via I/O port 205.

[00076] Storage controller 203 includes an event generator 224 that can issue events/message based on internal errors.

[00077] Port Monitor 216 tracks port statistics for all FC ports 217 on Fabric Switch 201. FC ports 217 are used to couple plural devices to switch 201. The statistics monitored including total words transmitted and received, CRC errors, link resets, invalid transmission words, loss of sync and excessive congestion and other link error conditions.

[00078] When a status value exceeds the threshold values configured/stored in Threshold Table 214, a control signal is sent from Port Monitor 216 to Performance Threshold Monitor 209 for further processing of the event. In one aspect of the present invention, a port is taken offline and/or a Performance Threshold Monitor 209 notifies host computer 202 of the threshold variation/violation.

[00079] Fabric Monitor 215 tracks the operation of Remote Fabric Switch 204 in a SAN. Fabric Monitor 215 issues keep-alive handshake messages through I/O Port C1 211 to other switches (for example, switch 204) in the SAN. Fabric Monitor 215 supports a plurality of instances of I/O Port C1 modules 211 to monitor a plurality of Remote Fabric Switch 204 in the SAN. The I/O Port C1 modules 211 may be based on Ethernet or Fibre Channel standards.

[00080] When messages/information cannot be exchanged with a given Remote Fabric Switch 204, Fabric Monitor 215 sends a control signal to Performance Threshold Monitor 209. Fabric Monitor 215 also receives Remote Fabric Switch events such as forced fail-over for I/O blade failure and for switches being taken offline for servicing. Fabric

Monitor 215 issues control signals to Performance Threshold Monitor 209 for further processing of these events.

[00081] In one aspect of the present invention, when a remote fabric switch 204 does not respond, the E_Port connecting to the remote fabric switch 204 is taken off-line. An event command may also be sent to host 202. In case of a partial failure, host computer 202 may take over a portion of the remote switch data path.

[00082] Performance Threshold Monitor 209 consolidates the threshold events and control signals sent by the monitoring agents and if the thresholds defined in Threshold Table 214 have been exceeded, a control signal is sent to Event Response Director 208 for further processing.

[00083] Event Response Director 208 uses the control signals sent by Performance Threshold Monitor 209 to select an event (routing and handling) as defined in Event Response Table 213. Based on the selected event routing and handling definition, Event Response Director 208 sends control signals to Local Event Handler 207 and External Event Generator 206.

[00084] Local Event Handler 207 performs control actions on the local Fabric Switch 201. This includes operations such as taking an E_Port down and forcing new inter-switch routing as well as shutting down the switch in an over-temperature situation.

[00085] External Event Generator 206 prepares and sends I/O PathGuard RSCNs (Registered State Change Notice) to registered Host Computers (202) through the I/O Port A1 interface 205. External Event Generator 206 supports a plurality of instances of I/O Port A1 modules 205 to notify a plurality of Host Computers 202. Host Computer I/O Path Failover Controller 222 receives the I/O PathGuard RSCN and uses this event to trigger a path failover whereby SCSI traffic is redirected to be sent over an alternative path.

[00086] At the completion of the path failover processing, Host Computer I/O Path Failover Controller 202 sends a control signal back to the Fabric Switch via I/O Port A2 221. External Event Controller 206 forwards Host Computer 202 response to Event Response Director 208 to complete the event handling.

[00087] Threshold Event Configuration module 220 updates the Event Response Table 213 and the Threshold Table 214 based on user requests using host computer 202. Tables 213 and 214 can be updated based on history of certain failures and any other reliability data that may be useful to predict failures.

[00088] Host Computer 202 may initiate diagnostic action via I/O Path Diagnostic Controller 223 by sending a control signal to the Fabric Switch Diagnostic Control module (not shown). The Diagnostic Control may initiate loop back diagnostics and supports diagnostic probes to the attached devices.

[00089] Figure 3 shows a process flow diagram for monitoring events in a SAN and then responding to events to avoid disruption in the SAN operation.

[00090] Turning in detail to Figure 3, in step S300, a fabric switch 201 monitors plural events occurring in connection with plural monitoring agents. For example, Chassis Monitor 218, as described above with respect to Figure 2, tracks the temperature of the switch chassis (not shown), the power supply status, and fan (not shown) speed; Nx_Port Event Monitor 212 receives events

generated by SAN targets such as Storage Controller 203; Port Monitor 216 tracks port statistics for all the FC ports 217 on Fabric Switch 201; and Fabric Monitor 215 tracks the operation of Remote Fabric Switch 204 in a SAN.

[00091] In step S301, the process compares an events status value with a threshold value for the event. For example, in the case of a Chassis Monitor 218, as described above with respect to Figure 2, when a status value exceeds the threshold value configured in Threshold Table 214, a control signal (this term as used in the specification includes commands and/or message packets) is sent from Chassis Monitor 218 to Performance Threshold Monitor 209 for further processing of the event. Other agents perform the same.

[00092] In step S302, the process consolidates the various control signals from plural monitoring agents and then processes the events. Performance Threshold Monitor 209 processes the events, as described above.

[00093] In one aspect of the present invention, various events are monitored in real time, which can result in the failure of certain SAN components/services. This allows a system to be

intelligent and retroactive, which avoids disruption of SANs.

[00094] Although the present invention has been described with reference to specific embodiments, these embodiments are illustrative only and not limiting. For example, the various modules shown in Figure 2 may be consolidated as a single software and/or hardware function to perform the various operations that have been described above. Many other applications and embodiments of the present invention will be apparent in light of this disclosure and the following claims.